IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 25, 2021, accepted March 2, 2021, date of publication March 4, 2021, date of current version April 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063944

Prediction of Second Primary Lung Cancer Patient's Survivability Based on Improved Eigenvector Centrality-Based Feature Selection

PENG LIU[1], (Member, IEEE), KEXIN JIN[2], YIPING JIAO[1], MUTIAN HE[3], AND SHUMIN FEI[1], (Member, IEEE)

[1]Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
[2]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA
[3]Department of Electrical Engineering and Control Science, Nanjing Tech University, Nanjing 210031, China

Corresponding author: Shumin Fei (smfei@seu.edu.cn)

ABSTRACT Modeling of second primary lung cancer (SPLC) patients' survival prediction has important theoretical significance and practical needs. Cancer survivability prediction may provide advice for better clinical decisions and personalized medicine. The Surveillance, Epidemiology, and End Results (SEER) program provides large data sets for analysis with machine learning methods. SPLC cases are identified and labeled from the SEER database; the data set is then preprocessed with improved eigenvector centrality-based feature selection (IECFS). The IECFS method utilizes interclass and intraclass dispersions and the ranking criteria. By adjusting the value of the α parameter and the number of features selected, the method achieves the best performance. The experiment is divided into five folds. This method yields a prediction accuracy of 90.998% for the five-year survivability that is higher than the original classification accuracy (89.16%) and the other state-of-the-art feature selection methods. For the three-year survivability, the proposed methods yields a prediction accuracy of 83.16%, slightly outperforming all of the compared methods. The method is effective and generalizable.

INDEX TERMS Second primary lung cancer, cancer survival prediction, improved eigenvector centrality-based feature selection.

I. INTRODUCTION

In the past 70 years, cancer prognosis improved markedly due to the promotion of cancer screening, development of medical technologies, and advances in supportive care. In the United States, the 5-year relative survival in 2016 was estimated to be 70%, twice as high as that in 1950s [1]. The number of cancer survivors in the United States will grow from 16.9 million in 2019 to a projected number of 22.1 million in 2030, accounting for 5% of the total population [2]. Due to the improved prognosis as well as aged population, multiple primary cancer (MPC) diagnoses for the same person are increasingly common. The risk of cancer survivors developing a second primary cancer was estimated to be 14% higher than that of the general population [3]. In the United States, one in five cancers diagnosed today occurs in an individual with a previous history of cancer [4]. On the other hand, MPC is much harder to treat due to a narrower range of options. For example, MPC patients may have previously received a maximum life-time dose of certain chemotherapy or the same part of the body may have undergone radiotherapy because of previous cancers [4], [5]. The prevalence and limited treatment options have made MPC an important issue for research, clinical treatment, and public health.

Second primary lung cancer (SPLC) has been the most common MPC, representing 25% of second primary malignancies [6]. In the Surveillance, Epidemiology, and End Results (SEER) program between 1992 and 2008, 1,450,837 non-pulmonary cancer survivors were identified, among whom 25,472 developed SPLC at a mean (standard deviation) follow-up of 5.7 (3.6) years. More than half (57%) of patients with SPLC died of the disease [6]. SPLC ranked only second to the same-site MPC in cases of prostate cancer and female breast cancer, the most common cancers among

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Hao Chen.

men and women, respectively [4]. The relatively high prevalence of SPLC is ascribed to the risk factors associated with MPC.

Most SPLC studies have focused on predicting the initial primary lung cancer patients' risks of developing SPLC. Other MPC survival prediction studies have been limited to genital MPCs. They applied statistical method to improve the prediction accuracy. Research on survival prediction of diagnosed SPLC patients has been lacking. Cancer survival rate prediction may provide guidance for better clinical decisions and personalized medicine. SPLC is the most common multiple primary cancer. However, few researchers have focused on the survival prediction of multiple primary cancer patients. Thus, prediction of the SPLC survival rate has become essential in cancer studies. Survivability often refers to the likelihood of a patient being alive after five years since the time of cancer diagnosis. It is an indicator in medical science for the evaluation of the treatment effectiveness. The method proposed in this paper predicts five-year and three-year survivability of SPLC patients. The novel IECFS method is applied to select features and to improve prediction accuracy.

The main contributions of this article are as follows:

- Identify and label the SPLC cases from SEER database and study the survival prediction of them;
- Apply improved eigenvector centrality feature selection;
- Compare prediction results with different amount of features and α;
- Compare prediction results with different feature selection methods; and
- Compare prediction results in different folds.

The average 5-year prediction accuracy of the proposed method is 90.998%, higher than the original method's accuracy (89.16%) and the results obtained by the compared FS methods. The 3-year prediction accuracy is 83.16%, higher than the original prediction accuracy (81.07%). The remainder of this article is organized as follows. Section 2 introduces the related works on SPLC, feature selection and the application of machine learning techniques to cancer survival prediction. Section 3 provides detailed methods and experimental procedures. Section 4 presents experimental results while section 5 presents the discussion of the results. Section 6 concludes the paper and presents possible future research directions.

II. RELATED WORK

MPC has been studied in epidemiology. The most important category of risk factors for MPC is life-style factors, such as tobacco smoking and alcohol consumption. According to nine SEER registries, an estimated one-third of MPC happened in tobacco- and alcohol-related sites between 1975 and 2000 [3], [4]. Tobacco smoking was defined as a Group 1 human carcinogen according to the International Agency for Research on Cancer (IARC) [7]. It is the leading risk factor for lung cancer at large [8]. Tobacco smoking is linked to approximately 80-90% of lung cancer deaths [9]. The other most important type of risk factor for MPC is prior cancer treatment, such as chemotherapy and radiotherapy. It was reported in nested case-control studies among the European and North American populations that larger numbers of chemotherapy cycles with alkylating agents elevated the risk of lung cancer among Hodgkin lymphoma survivors [10]. It was suggested that 8% of MPC is due to radiotherapy [11]. Prior chemotherapy was also observed to additively enhance the increased risk of second primary lung cancer by the previous radiotherapy among multiple types of cancer survivors [10], [12], [13].

The analysis of big data in health care and medical fields has immense potential for improving the quality of care, reducing medical waste, and reducing the burden of care [14]. Machine learning techniques have been widely applied to medical big data to predict outcomes [15], [16]–[18]. Liu *et al.* applied an improved clustering algorithm for sample cutting to improve training sample category representation capability. The experimental results indicated that the improved method improves the classification efficiency [19]. Ensemble learning methods that train a number of weak base learners and then combine their outputs are popular in medical prediction research [20]. Many researchers conducted their studies on cases collected from the SEER database. A Gaussian k-based naive Bayes (NB) classifier system was proposed by Kaviarasi *et al.* [21] to enhance the classification accuracy of the NB classifier and linear regression algorithm. They proposed an online gradient boosting learning with adaptive linear regressor and compared its performance with state-of-the-art machine learning algorithms.

Some researchers compared machine learning techniques with statistical methods to predict survivability for spinal ependymoma patients. They discovered that lower grade histology and greater extent of surgical resection were the key prognostic factors and concluded that therapeutic factors are associated with improved overall survival. Machine learning methods are generally better for prediction, but the data set was heterogeneous and complex with numerous missing values [22]. Several recently published papers on breast cancer survival prediction were analyzed together for application to stage-specific prediction tasks. Stage-specific prediction models and joint models were created and compared. It was concluded that data-driven knowledge obtained with machine learning methods must be subject to over-time validation prior to its clinical and professional application [23].

Roffo *et al.* proposed several feature selection methods. Reference [24] introduced an infinite feature selection method exploiting the convergence properties of power series of matrices. The Spearman's rank correlation coefficient and the standard deviation were combined and utilized. Replacing the standard deviation measure with dispersion criteria and applying the improved method to multiple primary cancers, we proposed a two-stage prediction method [25]. In 2017, Roffo *et al.* proposed a feature selection method via eigenvector centrality [26]. They built a graph to measure class separability based on mutual information and standard deviation.

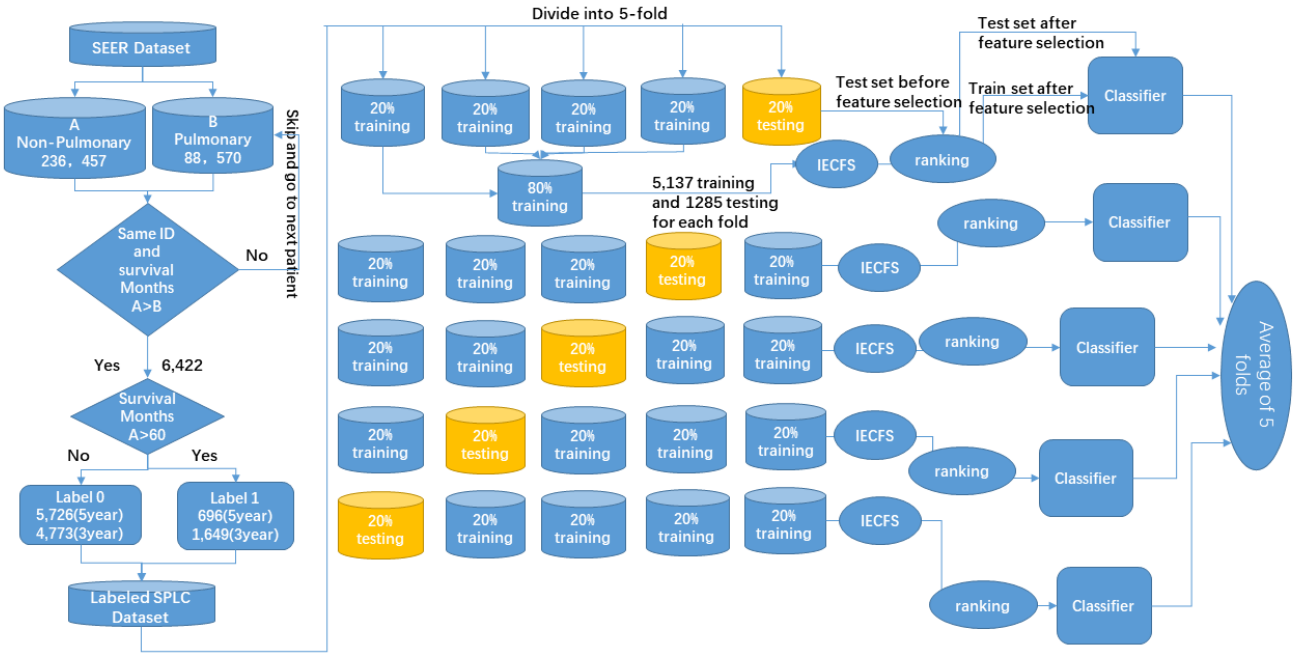


FIGURE 1. Flowchart of the proposed model.

Later, they proposed a new feature infinite latent feature selection method [27].

Zolbanin *et al.* used cancer data collected from the SEER database to create two comorbid datasets: one for breast and female genital cancers and another for prostate and urinary cancers. Several popular machine learning techniques were then applied to the resultant data sets to build predictive models. The results showed that the availability of more information about the comorbid conditions of patients improved the predictive power of the model [28]. In addition to Zolbanin's study, Naghizadeh *et al.* also investigated comorbid cancer patients focusing on data preprocessing, including feature selection and data cleaning. The feature selection procedure was performed by applying the least angle regression, least absolute shrinkage and selection operator, and stepwise regression algorithms. They compared the performance of four machine learning techniques for survivability prediction of prostate cancer. It was found that neural network outperformed decision tree, naive Bayes, and support vector machine learning. The accuracy was increased, and the error rate was decreased [29]. In this study, they investigated the survivability for female and male MPCs. However, they did not discuss the number of features selected in their article. In our previous research on the MPC patients' survivability, 150 features were selected. Currently there is no survival prediction model for patients with SPLC. Reference [30] estimated the trends in 5-year incidence of metachronous SPLC and stablished a risk prediction model to identify candidates with high SPLC risks. Reference [31] estimated the 10-year risk of developing second primary lung cancer (SPLC) among the survivors of initial primary lung cancer (IPLC). This paper will study feature selection through grid searching in the Results and Discussion Section.

III. MATERIALS AND METHODS

Cancer survivability prediction has been challenging due to the lack of publicly available large-scale medical data. SEER is an open-source database that provides de-identified, coded, and annotated information on cancers in the United States [6], [7]. The scale of data is large enough for analysis. To predict the 5-year survival rate among SPLC patients, non-pulmonary cancer survivors with lung cancer as the second primary lung cancer were selected from the SEER (Incidence0 SEER 18 Regs Research Data, Nov 2018 Sub) database. The lung cancer survivors were excluded from the study because of their relatively poor prognosis. (The lung cancer 5-year survival rate is 21.2%, lower than most of other cancer types [32].) The cancer diagnoses in SEER cancer registries were all by law verified clinically or microscopically, by a recognized medical practitioner [33]. The diagnosis of non-pulmonary cancers preceded the lung cancer diagnosis for each individual subject in this study.

Figure 1 is the flowchart of the proposed framework. The steps are as follows:

1) Collect the data on non-pulmonary cancers and pulmonary cancer from the SEER database;
2) Combine and label the data to create the SPLC data set, and determine the survival rate;
3) Divide the data into 5 folds of the same size, repeat (4)-(6) five times so that each portion has been used as the testing set;

4) Select optimal features for modeling according to IECFS;
5) Apply linear SVM as the classifier;
6) Record the predicted outcomes with the following error criteria: accuracy, sensitivity, specificity, NPV, and AUC and go to the next fold;
7) Calculate the average of the performance metrics; and
8) Compare the averaged metrics for different α values and different amount of features.

The case numbers are shown on the left side of Figure 1. After creating the SPLC data set, 6422 patients are divided into five sub-groups of similar sizes. In the first fold, the first four subgroups are used as the training set and the fifth is used as the testing set. After going through IECFS and SVM classifier, the test results are recorded. This process is repeated five times so that all of the subgroups are tested and recorded. Taking the average of the five folds, the final results are achieved. Adjusting α and the number of features for IECFS, the optimal results can be found.

A. DATA ACQUISITION

The clinical data are collected from the SEER database. The SEER program collects cancer data throughout the United States with the goal of reducing the cancer burden ultimately. SEER-Stat is a software developed to provide easy access to data analysis [6].

Non-pulmonary cancer cases are extracted from the SEER database first. Cases with 'positive histology', 'complete dates', 'active follow-up' are chosen, while cases with 'autopsy only', 'death certificate only', 'unknown cause of death', and 'unknown stage' are excluded. Benign and in situ cancer patients are excluded since they can be cured at a much greater chance and should be treated differently. Non-epithelial skin cancer patients are excluded due to its low mortality rate. Cases diagnosed after 2014 are excluded in order to predict the study participants' five-year survivability. Pulmonary cases are extracted from the SEER database to form a second data set. Most of the selection criteria are the same except that only pulmonary cancer patients are chosen.

The SEER database provides many attributes. Some of the attributes are similar to each other, while some unrelated to this research. Table 1 lists the key attributes selected. 26 features are selected in both the non-pulmonary data set and the pulmonary data set in this research. Researchers have previously studied the survival time prediction for lung cancer patients [34]. They selected 19 features from the SEER database. This study investigated the survivability prediction for SPLC patients. 18 features from [34] were selected and 8 more are kept. The only feature not selected is Radiation. This feature is no longer available in the database. The added 8 features include: patient ID, marital status, state-county, behavior, race, year of diagnosis, and sex. The patient ID feature is added to select the SPLC patients from the SEER database. Marital status, state-county, behavior, race, and sex have been shown to be related to the patients' prognosis [6]. The year of diagnosis is very important since cancer prognosis has been improved markedly and rapidly [1]. Most features are discrete and others are continuous. Table 1 includes the chosen features and a brief description of each feature. Discrete features are one-hot encoded and the continuous features remains unchanged prior to feature selection.

TABLE 1. Selected SEER attributes and their descriptors.

Feature Name	Description	Type
Patient ID	Patient ID, unique for each patient	discrete
Sex	Gender of patient	discrete
Age	Age of patient	numeric
Marital status	Marital status of patient at diagnosis	discrete
Grade	Grading and differentiation codes	discrete
State-county	Origin of the patient	discrete
Behavior	Behavior recode for analysis	discrete
Number of tumor	Total number of in situ/malignant tumors	numeric
Race	Race of Patient	discrete
Year of diagnosis	The year when patient was diagnosed	discrete
T	Derived AJCC T, 6th ed (2004-2015)	discrete
N	Derived AJCC N, 6th ed (2004-2015)	discrete
M	Derived AJCC M, 6th ed (2004-2015)	discrete
Stage	Derived AJCC stage group, 6th ed (2004-2015)	discrete
Primary Site	Location of the tumor	discrete
First malignant primary indicator	Based on SEER database record	discrete
Sequence number	Order of cancer occurrence with respect to other cancers	discrete
Lymph nodes	CS lymph nodes (2004-2015)	numeric
Histology recode	Broad grouping	discrete
Histologic Type ICD-O-3	Based on histologic type ICD-O-3.	discrete
Record number recode	Number of records	numeric
RX Summ-Scope Reg LN Sur (2003+)	Procedure of removal, biopsy, or aspiration of regional lymph nodes.	discrete
RX Summ–Surg Prim Site (1998+)	Procedures to remove or destroy tissue of the tumor.	discrete
Derived SS1977(2004-2015)	Derived "SEER Summary Stage 1977" from the CS algorithm	discrete
Tumor Size	Measurement of tumor size	discrete
Survival Months	Time between death and diagnosis	numeric

B. COMBINING DATA FOR SPLC CASES

Approximately 413,138 patients are identified with first incident cancers diagnosed before 2014 that meet the inclusion criteria. Then, 88,569 patients with lung cancer are included in the second data set. To find patients with SPLC, the records with the same IDs are extracted from the two sets of data. If lung cancer is diagnosed after the first incident cancer, the record is considered as a SPLC and is added to the final data set. Dropping patient IDs and survival months, and including marital status, sex, age, race, and state-county, the rest of the chosen features are kept for feature selection described in the next paragraph. Survival in months since the diagnosis of second primary lung cancer is considered to be the patient's survival time. The survival time tab is transferred to five-year and three-year survivability to be predicted. Patients who lived over 60 months are labeled as 1, while others are labeled as 0 in five-year survivability

prediction. In three-year survivability prediction, the 1 label is given to patients who lived over 36 months.

Approximately 6,422 cases remain after the selection process described above. Figures 2 and 3, and Table 2 display the distribution of the patients' survival time. About one-tenth of the patients lived more than 60 months since the diagnosis of the first incidence cancer. About one quarter of the patients lived more than 36 months.

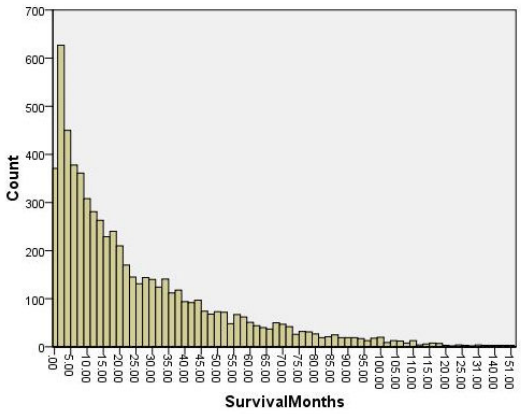


FIGURE 2. Histogram of patients' survival time.

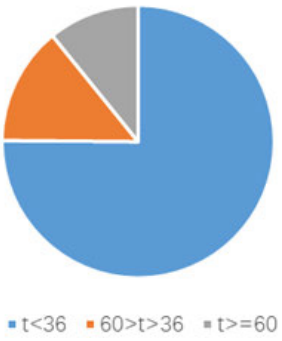


FIGURE 3. Patients' survival time distribution.

TABLE 2. Percentage of patients' survival time.

Survival Time	$t < 36$	$36 \leq t < 60$	$t \geq 60$
Number	4825	901	696
Percentage	75.13%	14.03%	10.84%

C. FEATURE SELECTION ACCORDING TO THE IMPROVED EIGENVECTOR CENTRALITY FEATURE SELECTION

Feature selection, also called feature subset selection, or attribute selection, is the process of selecting a subset from the feature group to improve model performance. In the application of machine learning, a large number of features is always present. Some features may be irrelevant to the label, and some may be dependent on each other. The irrelevant and redundant features may lead to lengthy training time, over-complicated model, and low generalization.

The original ECFS model jointly considers the variation of the features (maximum standard deviation over two features) and the relation of the two features to the class (mutual information). This method ranks each feature f_j according to the score s_j through calculation as the priority of each feature to be selected. In the actual construction of the model, n features can be selected from the top to the bottom by priority.

Specifically, given a training set F represented as $F = \{f^{(1)}, \ldots, f^{(n)}\}$, a nondirected fully connected graph $G = (V, E)$ can be built. The vertices V correspond to all features, while the edges E represent the pairwise relations between features. G can be represented as an adjacency matrix A whose elements a_{ij} $(1 \leq i, j \leq n)$ represent the pairwise relationship between features [24]. The elements are called pairwise probabilistic energy terms that are defined as:

$$a_{ij} = \varphi(f^i, f^j) \tag{1}$$

In ECFS, they use:

$$\varphi(f^i, f^j) = \alpha k + (1 - \alpha)\Sigma, \tag{2}$$

where k is the mutual information part and Σ is the maximum of the standard deviation of the two features.

Next, ECFS configure the priority for each feature by quantifying the path probability passing through a feature node. To measure the discriminative power of a single node, all possible paths that go through the node must be considered. γ_{ij} denotes a path of length l between nodes i and j through other features, we can then estimate the probability using Eqs. [3]-[5]:

$$P_\gamma = \prod_{k=0}^{l-1} a_{v_k, v_{k+1}} \tag{3}$$

To account for the energy of all possible paths of the length l, $P^l_{i,j}$ is defined as the set of all paths of l between i and j:

$$R_l(i, j) = \sum_{\gamma \in P^l_{i,j}} P_\gamma \tag{4}$$

which is equivalent to:

$$R_l = A^l, \tag{5}$$

It was proven that as l approaches a large number L, $A^l\mathbf{e}$ converges to v_0 [26]:

$$\lim_{l \to L} [A^l e] = v_0 \tag{6}$$

Therefore, ECFS can be realized by finding the eigenvalues and eigenvectors of matrix A:

$$det(\lambda I - A) = \begin{vmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{vmatrix} \tag{7}$$

The eigenvectors can be then calculated by solving Eq. (8):

$$(\lambda I - A)v = 0 \tag{8}$$

The absolute value of v_{ij} represents the score of the j^{th} node to the i^{th} node. The score of the j^{th} node can be calculated with equation (8). v_{ij} denotes the j^{th} element of the eigenvector v_i, $i = 1, \ldots, n$:

$$s_j = \sum_{i=1}^{n} |v_{ij}| \tag{9}$$

$$s = \begin{bmatrix} s_1 & s_2 & \cdots & s_n \end{bmatrix} \tag{10}$$

The original ECFS method use standard deviation and mutual information to measure the complexity of two features and their relationship. However, these two parts are not comprehensive. We introduce our IECFS method where the A matrix is different. The adjacency matrix A containing a_{ij} of the graph G is defined as:

$$A = \alpha D_b + (1 - \alpha) D_w \tag{11}$$

where D_b and D_w represent interclass dispersion and intraclass dispersion.

$$D_b = \sum_{i=1}^{c} p(C_i)(\bar{f}_i - \bar{f})(\bar{f}_i - \bar{f})^T \tag{12}$$

$$D_w = \sum_{i=1}^{c} p(C_i) \frac{1}{n_i} \sum_{j=1}^{n_j} (f_{ij} - \bar{f}_i)(f_{ij} - \bar{f}_i)^T \tag{13}$$

$$D_t = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{n_j} (f_{ij} - \bar{f}_i)(f_{ij} - \bar{f}_i)^T \tag{14}$$

D_b represents interclass dispersion, which is the difference between the samples in two classes. D_w is the intraclass dispersion, representing the variance of the samples in the same class. D_t is the total dispersion, equivalent to the sum of D_b and D_w. The dispersion between cases (D_t) reflects the average difference between the samples. Large intraclass dispersion (D_w) reflects large difference between different classes, while large interclass dispersion (D_b) reflects large variance between the samples in the same class. The intraclass and interclass dispersion represent the samples' separability in two aspects.

$\alpha \in [0, 1]$ is a hyperparameter that balances the importance of D_w and D_b. It can be adjusted to improve classification performance.

To optimize the results of IECFS, we need to find the best performing α and combination of features. The optimal combination was found through grid-searching. All of the accuracies and AUCs of $\alpha \in [0.1, 0.9]$ at the step of 0.05 and number of features $n \in [100, 150, 200]$ were calculated and are listed in Tables 3 and 4. The elements in matrix A represent the discriminative power when i^{th} and j^{th} elements are jointly considered.

TABLE 3. Average of 5 folds' accuracy and AUC for 5-year survivability with different α.

α	ACC	AUC	ACC	AUC	ACC	AUC
Feat	100		**150**		200	
0.1	0.9049	0.9149	0.9057	0.9134	0.9066	0.9108
0.15	0.9024	0.9159	0.9052	0.914	0.9072	0.9102
0.2	0.906	0.9162	0.9066	0.914	0.9063	0.9109
0.25	0.9029	0.9165	0.9052	0.915	0.9056	0.9093
0.3	0.9035	0.9161	0.9064	0.9143	0.9053	0.9095
0.35	0.9028	0.915	0.9069	0.9148	0.906	0.9099
0.4	0.9027	0.9154	0.9069	0.9146	0.9066	0.9097
0.45	0.9028	0.9153	0.9083	0.915	0.9067	0.9099
0.5	0.9033	0.915	0.908	0.9142	0.9066	0.9098
0.55	0.9041	.09149	0.908	0.9142	0.9069	0.9098
0.6	0.9041	0.9146	0.908	0.9142	0.9069	0.9099
0.65	0.9041	0.9146	0.9076	0.9142	0.9072	0.9101
0.7	0.9041	0.9146	0.9087	0.9144	0.9066	0.91
0.75	0.9041	0.9146	0.9092	0.914	0.9063	0.9097
0.8	0.9095	0.914	**0.91**	**0.9143**	0.9064	0.9099
0.85	0.9033	0.914	0.9092	0.914	0.9064	0.9104
0.9	0.9031	0.914	0.9095	0.9138	0.9064	0.9104

TABLE 4. Average of 5 folds' accuracy and AUC for 3-year survivability with different α.

α	ACC	AUC	ACC	AUC	ACC	AUC
Feat	**100**		150		200	
0.1	0.8301	0.8812	0.8264	0.8695	0.8275	0.8784
0.15	0.829	0.88	0.8276	0.8833	0.8286	0.8789
0.2	0.829	0.8798	0.8272	0.8807	0.8279	0.8787
0.25	0.83	0.8786	0.8278	0.881	0.8282	0.8795
0.3	0.8317	0.8811	0.829	0.881	0.8295	0.8793
0.35	0.831	0.881	**0.8317**	**0.8811**	0.8295	0.8796
0.4	0.8317	0.8811	0.8309	0.8813	0.8295	0.8794
0.45	0.8317	0.8811	0.8306	0.8813	0.8289	0.8793
0.5	0.8306	0.8809	0.8304	0.8814	0.8289	0.8793
0.55	0.8306	0.8809	0.8307	0.8812	0.8295	0.8793
0.6	0.8306	0.8809	0.8307	0.8812	0.8295	0.8795
0.65	0.8306	0.8809	0.8301	0.8816	0.8295	0.8795
0.7	0.8306	0.8809	0.8303	0.8816	0.8295	0.8795
0.75	0.8306	0.8809	0.8303	0.8816	0.8297	0.8794
0.8	0.8306	0.8809	0.8303	0.8816	0.8297	0.8794
0.85	0.8301	0.8808	0.831	0.8815	0.8297	0.8794
0.9	0.8301	0.8808	0.831	0.8815	0.8297	0.8794

D. EXPERIMENTAL PROCEDURES

The data are then randomly divided into five subgroups. Four out of the five subgroups are used as the training set and the rest of the data are used as the testing set. The experiment is repeated five times so that each subgroup has been used as the test data. 5138 and 1284 observations were included in the training and testing data sets, respectively. When combining the features of the two cancers, some features were the same. Excluding these features from the feature pool, 40 features were kept and transformed to one-hot encoded 1687-dimensional data of zeros and ones. The improved-ECFS reduced the data dimensionality. Different values of α were adopted for comparison. Linear SVM was adopted as the classifier. The compared feature selection methods in the classification stage are as follows: mutual information-based feature selection, pairwise correlation-based feature selection method, and the original ECFS [35].

E. PERFORMANCE METRICS

The classification accuracy is quantified as recognition accuracy, precision, and recall. The formulas are as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Specificity = \frac{TN}{TN + FP} \tag{16}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{17}$$

$$NPV = \frac{TN}{TN + FN} \tag{18}$$

True positive (TP) represents the patients who lived longer than 60 months and were predicted to do so, True negative (TN) represents the patients who did not survive up to 60 months and whose prediction was also negative. FP (false positive) is the number of patients who did not live up to 60 months but were predicted to be positive, and FN (false negative) is the number of patients whose labels are 1 but are predicted to be 0.

Accuracy is the ratio of all correctly predicted cases to the total sample. Specificity measures the ability of the classifier to predict negative cases and is the fraction of the correct negative predictions over all negative samples. The sensitivity measures the classifier's performance for positive cases and is the fraction of correct positive predictions over all positive samples. These two metrics are commonly applied to medical classifiers. NPV is the abbreviation for negative predictive value. It reflects the probability that a predicted negative is a true negative [36].

F. SIMULATION SETUP

This proposed method is implemented in MATLAB 2015b. The operating system is 64 bit windows 10. The RAM memory is 16 GB, and the processor is an Intel(R) Core(TM) i7-6700HQ CPU @2.60GHz 2.59 GHz. The compared feature selection methods can be found in [27].

The classifier adopted is support vector machine. The MATLAB fitcsvm function with default sequential minimal optimization (SMO) algorithm is adopted.

IV. RESULTS

In the feature selection step, each feature is assigned a score. The features are then ranked according to the score. The feature with the highest score ranks the first and the feature with the lowest score ranks last. The number of features are then selected according to the ranking. If N features are selected, then the top N features are selected.

Tables 3 and 4 include the grid-searching results for the best α and number of feature selections combination. The α values ranging from 0.1 to 0.9 are tested. The number of feature selections are 100, 150, and 200. The best combinations are marked in bold in the tables.

Tables 5 and 6 contain the confusion matrices of TP TN FP and FN, the performance metrics including accuracy, specificity, sensitivity, the negative predictive value (NPV), and the area under curve (AUC). The best metrics with the same amount of feature selection are also marked in bold. In addition to the proposed improved eigenvector centrality-based feature selection method (IECFS), the compared feature selection methods are mutual information-based feature selection (muteInf) [37], correlation-based feature selection (CFS), and the original eigenvector centrality-based feature selection (ECFS) [26].

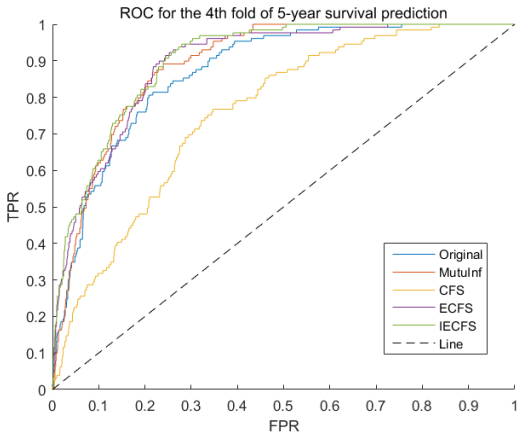


FIGURE 4. ROC for one of the five folds of the 5-year survival prediction.

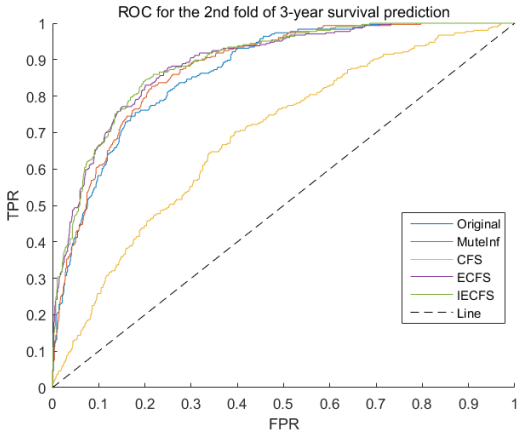


FIGURE 5. ROC for one of the five folds of the 3-year survival prediction.

Figures 4 and 5 show the ROC curves for one of the folds.

Figures 6 and 7 display the accuracy score for all five folds with 150 features. These two figures are plotted to show the consistency and the metabolizability of the proposed IECFS method.

V. DISCUSSION

Tables 3 and 4 display the average accuracies and AUCs for α and number of features combination. The best combination for the 5-year survivability prediction is $\alpha = 0.8$ $N = 150$. The best combination for the 3-year survivability prediction is $\alpha = 0.35$ $N = 150$. The best number of feature selection in the MPC patients' survival prediction is also 150 [25].

TABLE 5. Classification results for 5-year survivability.

Methods	TP	TN	FP	FN	Feat Sel	ACC	SP	SS	NPV	AUC
Original	64	1081.2	64	75.2	N.A.	89.16%	0.9441	0.4594	0.935	0.8888
MuteInf	9.2	1133.6	11.6	130	100	88.974%	0.9899	0.0677	0.8972	0.8873
CFS	0	1145.2	0	139.2	100	89.16%	**1**	0	0.8916	0.6611
ECFS	48	1113.4	31.8	91.2	100	**90.4072%**	0.9722	0.3448	0.9243	0.9145
IECFS	49.2	1112	33.2	90	100	**90.4072%**	0.9710	**0.3537**	**0.9251**	**0.9146**
MuteInf	27.6	1122.2	22.8	111.6	150	89.522%	0.98	0.20	0.9096	0.9093
CFS	0.2	1147.2	0.2	136.8	150	89.332%	**0.9998**	0.0014	0.8934	0.7803
ECFS	52.8	1106.6	38	86.4	150	90.314%	0.9666	0.3795	0.9276	0.9105
IECFS	57.4	1111.4	33.8	81.8	150	**90.998%**	0.9705	**0.413**	**0.9314**	**0.9143**
MuteInf	41	1116.2	29	98.2	200	89.976%	0.9747	0.2949	0.9192	0.9066
CFS	3.6	1141	4.2	135.6	200	89.12%	**0.9963**	0.0261	0.8938	0.82
ECFS	55.8	1106.4	38	86.4	200	90.488%	0.661	0.4002	0.9299	**0.9103**
IECFS	57.4	1106.8	38.4	81.8	200	**90.642%**	0.9665	**0.4119**	**0.9312**	0.9099
MuteInf	48.2	1107.2	38	91	250	89.958%	0.9668	0.3452	0.9242	0.9023
CFS	13.6	1135.8	9.4	125.6	250	89.49%	**0.9914**	0.0986	0.9007	0.8546
ECFS	57	1100.4	44.8	82.2	250	90.108%	0.9609	0.4909	0.9305	**0.9079**
IECFS	59.4	1104.6	40.6	79.8	250	**92.626%**	0.9646	**0.4266**	**0.9326**	0.9062

TABLE 6. Classification results for 3-year survivability.

Methods	TP	TN	FP	FN	Feat Sel	ACC	SP	SS	NPV	AUC
Original	200.6	840.6	114	129.2	N.A.	81.07%	0.8806	0.6081	0.8667	0.8595
MuteInf	178.2	837.8	116.8	151.6	100	79.103%	0.8777	0.5413	0.8557	0.7501
CFS	55.8	910	44.6	274	100	75.19%	**0.9532**	0.1686	0.7695	0.7021
ECFS	206	860.4	94.2	123.8	100	83.03%	0.9013	0.6251	0.8741	**0.8822**
IECFS	208.6	858.8	95.8	121.2	100	**83.11%**	0.8993	**0.6331**	**0.8763**	0.8810
MuteInf	219.8	827.8	126.8	110	150	81.56%	0.8672	**0.6677**	**0.8828**	0.8776
CFS	104.4	879.8	74.8	225.4	150	76.63%	**0.9218**	0.3196	0.7968	0.7652
ECFS	205.4	858	96.6	124.4	150	82.79%	0.8987	0.6233	0.8733	0.8793
IECFS	210.6	857.6	97	119.2	150	**83.16%**	0.8984	0.6393	0.878	**0.8812**
MuteInf	212	844.4	110.2	117.8	200	82.25%	0.8845	**0.6435**	**0.8775**	0.878
CFS	130.2	870.2	84.4	199.6	200	77.89%	**0.9117**	0.3982	0.8152	0.7948
ECFS	207.6	854.4	99	122.2	200	82.76%	0.8962	0.6294	0.8748	0.8787
IECFS	204.8	860.6	94	125	200	**82.95%**	0.9015	0.6214	**0.8731**	**0.8797**
MuteInf	208.4	850.2	104.4	121.4	250	82.42%	0.8906	**0.6325**	**0.875**	0.8708
CFS	171.8	855.2	99.4	158	250	79.96%	0.8957	0.5217	0.8443	0.8379
ECFS	204.4	854.6	100	125.4	250	82.45%	0.8952	0.6197	0.872	0.8757
IECFS	203	1104.6	94.8	126.8	250	**82.75%**	**0.9006**	0.6157	0.8714	**0.8766**

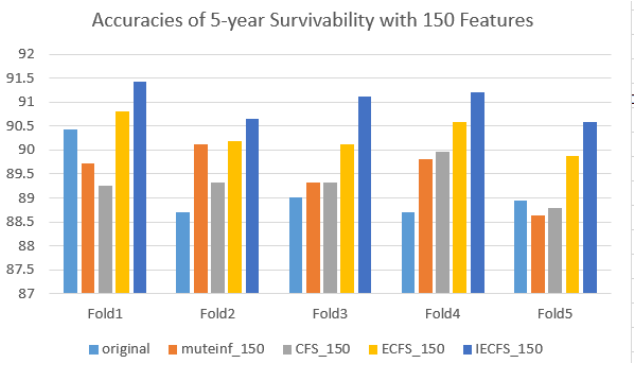


FIGURE 6. 5-year survivability's prediction accuracies of the five folds.

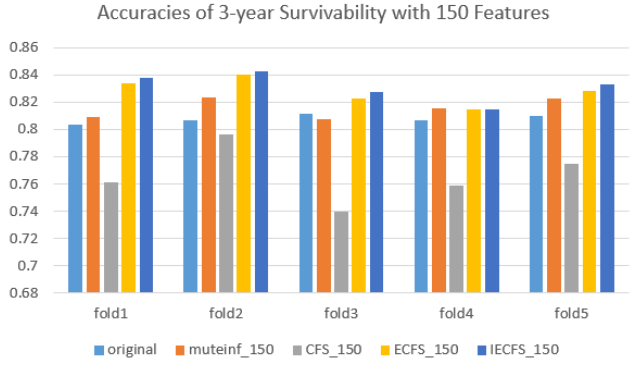


FIGURE 7. 3-year survivability's prediction accuracies of the five folds.

The AUC values do not match their accuracy scores. In some cases, the combination with high accuracy has low AUCs.

After selecting the best α and N, the performance of different feature selection methods are compared in Tables 5 and 6. The proposed IECFS method has the best accuracy scores among all of the feature selections. The improvement is not significant. This is caused by the five-fold averaging process.

The results for all five folds are recorded and averaged for the comparison. The IECFS has the best metrics except for the specificity. CFS has the poorest prediction but the best specificity. The three-year survival prediction also proves that the IECFS is the most suitable feature selection method. Interestingly, when 100 features are selected, the ECFS method achieves very good predicting result that is only slightly worse than that of the IECFS. The improvement of the three-year survivability is not significant but it is consistent in every fold. The proposed IECFS method achieves the best accuracy score in all folds. This proves the robustness of our model.

The comparison of the feature selection methods shows that the mutual-information-based feature selection and the correlation-based feature selection methods have poorer prediction outcome. However, when the number of the selected features increases, their performances improve. When only 100 features are selected, the MuteInf method has worse performance than the original prediction. As the number of features increase, it performs better than the original method. CFS's performance also improves with increasing number of features. However, when 250 features are selected, it is still worse than the original method. The ECFS and IECFS methods are different. In 3-year prediction, when 100 features are selected, these methods have the highest accuracy score when different number of features are selected. In 5-year prediction, when 250 features are selected, they have the poorest accuracy when different numbers of features are selected. This is caused by their ranking criteria.

Figures 4 and 5 shows the ROC curves for one of the five folds for three-year and five-year survival prediction. Three of the curves are close to each other. An examination of the data presented in Tables 5 and 6 shows that the AUC values of mutual-information-based feature selection, ECFS, and IECFS are close to each other.

As mentioned in the previous sections, the test is five-folded. The average performance metrics are calculated and compared. The five folds of survivability prediction are also plotted in Figures 6 and 7. For both Figures, 150 features are selected. Each fold has a different test set. Comparing Figure 6 with Figure 7, we observe that the IECFS displays a larger improvement of the prediction for five-year's survivability in all folds. The three-year's survivability prediction has a smaller improvement but it outperforms other methods in all folds.

VI. CONCLUSION

SPLC is the most common MPC. Predicting the survivability of SPLC patients can help the doctors, patients, and families. However, few researchers have studied the survival prediction of MPC patients. Thus, the prediction of SPLC survival rate has become essential in cancer studies. In this research, SPLC cases were identified and labeled to study survival prediction. The proposed IECFS method outperforms the state-of-the-art feature selection methods.

The IECFS method outperforms the compared methods in all five folds, proving that the IECFS method is robust and generalizable. The improvement of the IECFS method over the original ECFS method is moderate, but consistent. The IECFS method outperforms the ECFS method for a wide range of numbers of features.

This study focused on SPLC and proposed a novel IECFS method. We did not apply any data balancing method or data cleaning method because these methods introduce randomness into the prediction. In future work, we will consider utilizing the methods. Feature selection can be further improved by jointly considering multiple feature selection methods through statistical voting. In the future, we may also study other MPCs' survivability. We may also study the risk of developing MPCs after the initial primary cancers.

REFERENCES

[1] N. Howlader, A. M. Noone, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin, Eds., "SEER cancer statistics review, 1975–2016," Nat. Cancer Inst., Bethesda, MD, USA, Apr. 2019. [Online]. Available: https://seer.cancer.gov/csr/1975_2016/

[2] K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowland, K. D. Stein, R. Alteri, and A. Jemal, "Cancer treatment and survivorship statistics, 2016," *CA, A Cancer J. Clinicians*, vol. 66, no. 4, pp. 271–289, Jul. 2016.

[3] R. E. Curtis, "New malignancies among cancer survivors: SEER cancer registries," U.S. Dept. Health Hum. Services, Nat. Inst. Health, Tech. Rep. 1973-2000, 2006.

[4] M. Thun, M. S. Linet, J. R. Cerhan, C. A. Haiman, and D. Schottenfeld, *Cancer Epidemiology and Prevention*. London, U.K.: Oxford Univ. Press, 2017.

[5] S. E. Lipshultz, M. J. Adams, S. D. Colan, L. S. Constine, E. H. Herman, D. T. Hsu, M. M. Hudson, L. C. Kremer, D. C. Landy, T. L. Miller, K. C. Oeffinger, D. N. Rosenthal, C. A. Sable, S. E. Sallan, G. K. Singh, J. Steinberger, T. R. Cochran, and J. D. Wilkinson, "Long-term cardiovascular toxicity in children, adolescents, and young adults who receive cancer therapy: Pathophysiology, course, monitoring, management, prevention, and research directions: A scientific statement from the American heart association," *Circulation*, vol. 128, no. 17, pp. 1927–1995, Oct. 2013.

[6] N. M. Donin, L. Kwan, A. T. Lenis, A. Drakaki, and K. Chamie, "Second primary lung cancer in united states cancer survivors, 1992–2008," *Cancer Causes Control*, vol. 30, no. 5, pp. 465–475, May 2019.

[7] T. Smoking, "IARC monographs on the evaluation of carcinogenic risks to humans 1986," in *Proc. IARC*, Lyon, France, 1986, pp. 1–1452.

[8] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[9] *Centers for Disease Control Control, and Prevention. 2019. What are the Risk factors for Lung Cancer?*, Division Cancer Prevention, Atlanta, GA, USA, Sep. 2019.

[10] E. S. Gilbert, M. Stovall, M. Gospodarowicz, F. E. van Leeuwen, M. Andersson, B. Glimelius, T. Joensuu, C. F. Lynch, R. E. Curtis, E. Holowaty, H. Storm, E. Pukkala, M. B. van't Veer, J. F. Fraumeni, J. D. Boice, E. A. Clarke, and L. B. Travis, "Lung cancer after treatment for Hodgkin's disease: Focus on radiation effects," *Radiat. Res.*, vol. 159, no. 2, pp. 161–173, Feb. 2003.

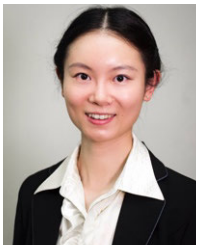
[11] A. B. de Gonzalez, R. E. Curtis, S. F. Kry, E. Gilbert, S. Lamart, C. D. Berg, M. Stovall, and E. Ron, "Proportion of second cancers attributable to radiotherapy treatment in adults: A cohort study in the US SEER cancer registries," *Lancet Oncol.*, vol. 12, no. 4, pp. 353–360, Apr. 2011.

[12] P. D. Inskip, M. Stovall, and J. T. Flannery, "Lung cancer risk and radiation dose among women treated for breast cancer," *JNCI J. Nat. Cancer Inst.*, vol. 86, no. 13, pp. 983–988, Jul. 1994.

[13] T. Grantzau, M. S. Thomsen, M. Væth, and J. Overgaard, "Risk of second primary lung cancer in women after radiotherapy for breast cancer," *Radiotherapy Oncol.*, vol. 111, no. 3, pp. 366–373, Jun. 2014.

[14] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," *Int. J. Med. Informat.*, vol. 114, pp. 57–65, Jun. 2018.
[15] Y.-Q. Liu, C. Wang, and L. Zhang, "Decision tree based predictive models for breast cancer survivability on imbalanced data," in *Proc. 3rd Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2009, pp. 1–4.
[16] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Breast cancer survivability via adaboost algorithms," in *Proc. 2nd Australas. Workshop Health Data Knowl. Manage.*, vol. 80. Darlinghurst, NSW, Australia: Australian Computer Society, 2008, pp. 55–64.
[17] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, "Robust predictive model for evaluating breast cancer survivability," *Eng. Appl. Artif. Intell.*, vol. 26, no. 9, pp. 2194–2205, Oct. 2013.
[18] L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou, and Y. Liu, "Reliable Parkinson's disease detection by analyzing handwritten drawings: Construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model," *IEEE Access*, vol. 7, pp. 116480–116489, 2019.
[19] H. Liu, S. Liu, and Z. Su, "Sample cutting and weighting method in text classification based on position," *Comput. Eng. Appl.*, vol. 2, no. 27, p. 2, 2015.
[20] L. Li, Q. Hu, X. Wu, and D. Yu, "Exploration of classification confidence in ensemble learning," *Pattern Recognit.*, vol. 47, no. 9, pp. 3120–3131, Sep. 2014.
[21] R. Kaviarasi, "Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed Gaussian classifier system," *J. Med. Syst.*, vol. 43, no. 7, p. 201, Jul. 2019.
[22] S. M. Ryu, S.-H. Lee, E.-S. Kim, and W. Eoh, "Predicting survival of patients with spinal ependymoma using machine learning algorithms with the SEER database," *World Neurosurgery*, vol. 124, pp. e331–e339, Apr. 2019.
[23] R. Kleinlein and D. Riaño, "Persistence of data-driven knowledge to predict breast cancer survival," *Int. J. Med. Informat.*, vol. 129, pp. 303–311, Sep. 2019.
[24] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4202–4210.
[25] P. Liu and S. Fei, "Two-stage prediction of comorbid cancer patient survivability based on improved infinite feature selection," *IEEE Access*, vol. 8, pp. 169559–169567, Aug. 2020.
[26] G. Roffo and S. Melzi, "Features selection via eigenvector centrality," in *Proc. New Frontiers Mining Complex Patterns (NFMCP)*, Oct. 2016, pp. 1–12.
[27] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, "Infinite latent feature selection: A probabilistic latent graph-based ranking approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1398–1406.
[28] H. M. Zolbanin, D. Delen, and A. H. Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decis. Support Syst.*, vol. 74, pp. 150–161, Jun. 2015.
[29] M. Naghizadeh and N. Habibi, "A model to predict the survivability of cancer comorbidity through ensemble learning approach," *Expert Syst.*, vol. 36, no. 3, Jun. 2019, Art. no. e12392.
[30] Z. G. Hu, W. X. Li, Y. S. Ruan, and F. J. Zeng, "Incidence trends and risk prediction nomogram of metachronous second primary lung cancer in lung cancer survivors," *PLoS ONE*, vol. 13, no. 12, Dec. 2018, Art. no. e0209002.
[31] S. S. Han, G. A. Rivera, M. C. Tammemägi, S. K. Plevritis, S. L. Gomez, I. Cheng, and H. A. Wakelee, "Risk stratification for second primary lung cancer," *J. Clin. Oncol.*, vol. 35, no. 25, p. 2893, 2017.
[32] C. Allemani *et al.*, "Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries," *The Lancet*, vol. 391, no. 10125, pp. 1023–1075, Mar. 2018.
[33] J. Ruhl, M. Adamo, and L. Dickie, "SEER program coding staging manual 2018," U.S. Dept. Health Hum. Services Nat. Inst. Health Nat. Cancer Inst., Nat. Cancer Inst., Bethesda, MD, USA, Tech. Rep. 20892, 2018.
[34] C. M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. de Carlo, J. A. Bartholomai, R. N. Balgemann, V. H. van Berkel, and H. B. Frieboes, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *Int. J. Med. Informat.*, vol. 108, pp. 1–8, Dec. 2017.
[35] Giorgio. *Feature Selection Library*. MATLAB Central File Exchange. Accessed: Jun. 24, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/56937-feature-selection-library
[36] H. Wang and H. Zheng, *Negative Predictive Value*. New York, NY, USA: Springer, 2013.
[37] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

PENG LIU (Member, IEEE) was born in 1990. She received the bachelor's degree from the Swanson School of Engineering, University of Pittsburgh, in 2012, and the master's degree from the School of Electrical Engineering, University of Washington, in 2014. She is currently pursuing the Ph.D. degree with Southeast University. Her research interests include data mining, medical data analysis, medical image analysis, machine learning algorithms, and deep learning.



KEXIN JIN is currently pursuing the Ph.D. degree with the Department of Epidemiology, University of California, Los Angeles (UCLA), CA, USA. She has enriched experience in the design and analysis of causal inference studies, such as clinical trials and real-world evidence, late phase clinical trials, relational medical claim databases, international surveys, case-controls, and cohort studies. She has published two first-author research articles in the *International Journal of Cancer* and *Translational Oncology*. Her research interests include causal inference in lung and colorectal cancers etiology, national policy, and health outcomes. She is currently an Active Member of the American Society of Preventive Oncology and the International Lung Cancer Consortium. She was awarded the Conrad N. Hilton Scholar at UCLA.

YIPING JIAO was born in 1990. He received the bachelor's degree in automation from Jiangnan University and the master's degree in automation from Southeast University, where he is currently pursuing the Ph.D. degree in medical image analysis and digital pathology.

MUTIAN HE was the Second Host In Charge of the Jiangsu-Level Project "Intelligent Double Truck Express Delivery System" (SRTP), from 2019 to 2020. He is currently a Junior with the Measurement and Control Technology and Instrumentation Program, Nanjing Tech University, Nanjing, Jiangsu, China. His research interests include computer vision, deep learning, image processing, and embedded system. His awards and honors include the First Prize of The 15th National College Students' Smart Car Competition, the First Prize of RoboMaster2020 (RMUT), the First Prize of the 10th Jiangsu College Students' Robots Competition, and the First Class Scholarship.

SHUMIN FEI (Member, IEEE) was born in 1961. He received the Ph.D. degree from Beihang University, Beijing, China, in 1995. From 1995 to1997, he did Postdoctoral Research with the Research Institute of Automation, Southeast University, Nanjing, China, where he is currently a Professor. His research interests include the analysis and synthesis of nonlinear systems, robust control, adaptive control, and the analysis and synthesis of time delay systems.